Evaluation: Principles and Starting Points

John Killeen, Senior Fellow, National Institute for Careers Education and Counselling

In summer 2003, the Guidance Council (with funding from the Department for Education and Skills) commissioned a report from NICEC to inform the early work of the new National Guidance Research Forum. The aim was to 'map' what research had already been done and who were the main players, and also what could be learned from existing research so that new work could build on what was known. With only two months to produce a report that could have taken a year or two, the editors addressed the challenge by suggesting a structure which the Forum might want to sustain, showing active research and funding agencies as well as identifying existing knowledge and good practice in individual areas of interest. NICEC members with expertise in particular fields were invited to provide statements on relevant work, within a common framework. John Killeen provided two of these sections (as well as other parts of the full report), and the first is reprinted here. His contributions to the report (Hawthorn, Killeen, Kidd & Watts, 2003) represented his last published work. This section in particular reflects many of the qualities which his colleagues and readers so valued. It is reprinted here with permission from the DfES and the Guidance Council.

I. What is it?

An extremely high proportion of all of the research done into guidance is 'evaluation', in the sense that it attempts to answer questions bearing upon effectiveness. But although there are quite a few guidance evaluators, their activities are so disparate that they do not form a coherent group with a common approach or shared literature. Moreover, many of them do not think of themselves in this way, despite the fact that at least part of the research they do is classifiable as evaluation.

Part of the reason for diversity is that guidance evaluation is conducted by people with very different kinds of expertise. In the UK, some of it is by *guidance practitioners* who are more or less consciously 'reflective practitioners', but not counselling psychologists. Some is by *expert ex-practitioners* whose style owes more than a little to the tradition of inspection in the public services. Quality standards are likely to have been written by similar kinds of people and tend to harmonise their criteria of judgement. In each case the fundamental working assumption is likely to be that guidance is a good thing (just as an OFSTED inspector might assume the value of education), but that it can be done more or less well, appropriately, and so on. These evaluators may use data collection techniques characteristic of social research, such as the 'semi-structured interview' and 'structured questionnaire', but few would claim to be expert in research methodology.

Ignoring the middle ground, one may characterise many others as having *research* expertise at the centre of what they do, to which may be added a research-based, rather than a practice-based, understanding of the world of employment, education, training, and guidance. Some claim expertise in 'qualitative'[1] techniques, and others in 'quantitative'[2] ones, whereas most research organisations of any size are able to combine both approaches in particular studies. The most difficult aspect of evaluation methodology is, however, the application of causal and statistical inference to the task of testing and quantifying effectiveness. People steeped in these methods take it as axiomatic that guidance is only a good thing if we can convincingly reject the null hypothesis that it is not. It is, therefore, difficult to understate the differences in mentality and the possibilities for mutual misunderstanding which exist amongst 'evaluators'.

In addition, various schisms occur, particularly in academia, so that, for example, 'top-down' evaluation, or the application of causal and statistical inference, are rejected, or through the insistence that evaluation should really be a place to do critical sociology, constructivist psychology, or whatever else it is that the person in question would rather do. However, the influence of such currents of thought on evaluation is not great.

[1] The term 'qualitative' is often used simply to mean 'non-statistical', although counts or other indications of magnitude (e.g. 'most', 'about half', 'few') may be reported. In routine social research practice, 'qualitative research' tends to mean case studies and the study of small samples. In this context, 'qualitative data' usually consists of the text of collected documents, records of observations and transcripts of 'open' questioning. Analysis consists of extracting meaning from text. Computer programs exist to aid this, but often go unused. The analytical process must usually be taken on trust. Not to be confused with what statisticians call 'qualitative data' (see note 2).

[2] In the social sciences, 'quantitative research' is that which involves measuring variables, usually in sizeable samples, for the purpose of subsequent statistical analysis. This can be in the context of an experiment, survey, etc. The variables can include what statisticians call 'qualitative data' (e.g. attributes such as male/female or experimental/control subject; or ordered attributes such as job satisfaction).

In this section we shall consider research associated with 'summative' evaluation, which examines the success of an intervention, initiative or organisation in achieving its goals, and any unintended benefits or dis-benefits it brings about.

Future reviews could focus in more depth on:

- Formative evaluation, or, evaluation intended to provide feedback whilst an intervention or initiative is under development or an organisation finds ways to achieve its goals.

- Evaluation of quality through the use of 'standards', which is to say assessment of the extent to which an intervention, practitioner, initiative or organisation meets good practice criteria.

Summative evaluation examines the success of an intervention or organisation in achieving its goals, whereas formative evaluation provides feedback whilst the means of doing so are developed. In practice, evaluation can involve a little of each.

At the heart of summative evaluation in guidance lies the intention to measure the *effects* or *impact* of interventions, or of the funding programmes through which they are delivered. There are two broad approaches to this task.

The first may be called 'interpretative'. Characteristically, it:

- is conducted and reported in relatively short time-scales;

- may be wholly 'qualitative' in character (e.g. reliant on small-sample, semi-structured interview and focus group data) or may gather 'quantitative' data from participants (e.g. from structured questionnaires completed by sizeable samples of service users);

- often occurs at the pilot stage of a new programme or as a small study for local accountability purposes;

- may also include a formative element;

- often gives at least as much attention to processes as outcomes;

- reports participant judgements about impact and makes more or less forensic use of participants' perceptions in order to form its own judgements about impact.

Thus, satisfaction ratings may be elicited and guidance clients may be asked to *attribute* effects to their guidance. Other participants and stakeholders, including practitioners, may also be asked to make these judgements, and evaluators may use information taken, ultimately, from participants' perceptions to construct critical judgements of their own.

EXAMPLE: 'INTERPRETATIVE' EVALUATION (ROLFE, 2000)

In 1997/98, DfEE funded 23 pilot projects intended to help young people in school to use labour market information (LMI). This was an evaluation of four pilot projects which were developing means to make customised LMI more accessible to young people, their advisers, teachers and parents. It was, therefore, partially formative in intent, consisting both of 'process' and 'impact' evaluation.

Impact was assessed in three localities, by holding group discussions and conducting depth interviews with Year 10 and Year 11 pupils, and by holding meetings and conducting interviews with participant teachers, careers advisers and others. These obtained self-report and opinion data, and looked for other evidence of impact.

It was reported, *inter alia*, that teachers and advisers had become more aware of LMI, that teachers needed LMI training and that a shortage of materials for special needs pupils had been made good. However, assessment of a key criterion, 'improved decision making', proved problematic. This was, in part, because young people learned how to access LMI, but made limited use of it outside classes.

Evaluation of this sort can provide useful management information and help practitioners reflect on their work. It can tell us whether clients and other participants believed they were helped by an intervention, and offer a *prima facie* case that it had particular effects. It can even provide evidence of how guidance helped particular people towards particular goals, as surely as one might establish in a court of law that the 9.30 from Euston took the defendant to Glasgow. But claims to identify impact in this way are insufficient, judged against the criterion of *additionality*, since there are other ways to reach Glasgow.

The second major form of guidance evaluation reserves the terms 'effect' and 'impact' for *changed probabilities of outcomes* or *changed probabilities of outcomes of given magnitudes*. This is often called 'quantitative' evaluation, in the sense that it seeks to quantify effects or impact defined in this way.

In order to measure effects or impact in this sense, it is necessary to compare the outcome for people exposed to guidance, or to a particular form of guidance, or to delivery under a particular funding programme ('treated'[3] subjects) to what the outcome *would have been*, if they had either remained untreated, or had been treated in another way. These

[3] Although the term 'treatment' comes from medical research, it is now used in a generic manner.

alternative futures clearly cannot be observed and must be estimated. The estimated outcome in the absence of treatment is called the 'counterfactual' ('counter' to what actually happened). Most of us are familiar with this in the form of a *control group outcome*; but, where this is not available, other ways of specifying the counterfactual are needed. The more general position is that we need to avoid doing this in a way that biases the estimate of effectiveness *within* the study.

Sometimes the intention is merely to demonstrate that a guidance intervention or programme *can* work. But we may want to *generalise*, which is to use the effects measured in a study to obtain estimates of effectiveness 'out there in the real world'. In principle, this can be divided into three parts: first, we may estimate the effect on currently treated subjects who are *like those in the study*[4]; second, we may estimate the effect on *all* current users or participants[5]; and third, we may estimate how much currently *eligible, but untreated,* people would benefit, if they were to participate[6]. These can all differ, because the people we study may not be representative of current participants, and current participants may differ from 'eligible but untreated' people.

There are two main branches of summative, quantitative guidance evaluation, which tend to get their counterfactuals in different ways. The first is composed of experimental trials which are methodologically similar to medical trials. In practice, this method is applied as follows:

- it is usually reserved to the study of very tightly defined guidance methods - often innovations;
- samples are usually small;
- outcomes tend to be assessed only over the short term;
- these are commonly learning and psychological outcomes rather than effects on careers.

Random assignment to the treatment and to a control (untreated) group, and/or to a group given an alternative treatment, is the paradigm method. This is because randomly assigned samples from the same pool are free of sample selection bias with respect to one another. Hence, we are on reasonably firm ground when we conclude that there was or was not an effect, or that there was a given size of effect, *within* any given study. However, guidance experimentalists are seldom concerned to generalise and to estimate effects in the world beyond themselves.

WHAT IS A RANDOM SAMPLE?

When every member of the intended study population is given an equal probability of entering a sample, this is called a 'simple random sample'. 'Random' does not mean 'casual': selection at random generally means that the whole population under consideration must be listed and every sample member selected from it by a specially-devised, lottery-like, 'random' method, so as to give each member of the population an equal probability of being drawn. A random sample is, by definition, unbiased. On the other hand, an 'opportunity sample' is one assembled on the basis of convenience (e.g. when all of the students in a class are treated as a sample of the student population, or when twenty people found in the street are treated as a sample of 'the population'). An opportunity sample is not a random sample and one can never be sure that it is unbiased.

In a randomised controlled trial, two samples are drawn from the same pool, one to be 'treated' (the 'experimental sample') and the other to act as an 'untreated', control sample. People are drawn so that they have an equal probability of entering either sample: they are 'randomly assigned', and 'sample selection bias' is avoided *within* the trial. But what of the pool from which they are taken? Perhaps the pool is a random sample of the relevant population? If so, the experimental and control samples are also random samples of, and unbiased representations of, that population. But perhaps the pool is an 'opportunity' pool (e.g. the students of a single class, when the population of interest is all students of this type)? Neither sample is randomly drawn from the whole, relevant client population. Generalisation of results beyond the experiment, to clients in general, becomes problematic. Most 'randomised controlled trials' of guidance are like this.

[4] Called the 'local average treatment effect' - the effect on the population of which those in the study are an unbiased sample. See Imbens & Angrist (1994).

[5] Called the 'average treatment effect on the treated'.

[6] Called the 'average treatment effect on the untreated'.

EXAMPLE: RANDOM ASSIGNMENT TRIAL (AUSTIN & GRANT, 1981)

The authors looked at the effect of interview skill training on 60 first-generation college students who were intending to seek work. Students were randomly assigned to groups of ten which received:

(a) didactic instruction;

(b) (a) + mock interview;

(c) (a) + professional videotape;

(d) (a) + self-video;

(e) (d) with feedback.

There was also a no-treatment control group.

The outcome measure was a judge-rated simulated job interview score.

All the treatments were found equally effective, yielding significant gains in mean scores relative to the no-treatment controls. Note the very small samples and the way in which everything hinges on just ten control subjects. To be statistically significant, differences have to be quite large when such tiny samples are compared. But also note the absence of an explicit 'placebo' treatment.

Of course, studies designed to be randomised trials may not actually achieve their ends. The most common problem is sample attrition, where individuals drop out, possibly for a reason relevant to the study's parameters. This can lead to *attrition bias*, which is a form of *sample selection bias*. Attrition is very common. For example, Whiston *et al.* (1998) assessed 70% of recent US trials to have 'attrition problems' above the mid-point of their rating scale.

EXAMPLE: NON EQUIVALENT GROUPS DESIGN (LENT, LARKIN & HASEGAWA, 1986)

This evaluation of a focused-interest career course for science and engineering undergraduates involved 54 science and engineering students and ten controls. The controls were students who had withdrawn from the course prior to the first session. The course offered ten sessions, each lasting 1 hour 45 minutes. Students were tested before and after the course, and the outcome measures were:

Decidedness;

Career Development Survey (CDS) scores for Self-Knowledge (interests values, skills) and Knowledge of Information;

Career information-seeking behaviour.

All tested positive; that is, the course was concluded to be effective. But the control group was self-selected, being made up of people who decided not to do the course. So it was not possible to assume that samples started out at the same level. When this is so, 'before-and-after' testing is of special importance. One looks for significantly larger *gains* in the experimental sample than in the control sample.

Studies which *look like* randomised trials are sometimes done by treating one pre-existing group, such as a school class or the clients of a particular agency, and using another pre-existing group as the 'control group'. Taken on their own, studies of this sort *are not an adequate substitute* for randomised trials because they are subject to sample selection bias. But providing that *groups* can be randomly assigned to the treatment and control conditions, and more groups can be added, studies like this can *become* adequate as the sample of *groups* increases in size. Although sufficiently sizeable single studies of this sort are not actually done, all is not lost, since the results of many small studies can be pooled (together with those of randomised trials). This pooling of results is called 'meta-analysis'.

The second main type of quantitative summative research is non-experimental and *ex post facto*. Much UK guidance evaluation which involves comparison to the counterfactual is of this type. In general, these studies:

- are into provision or programmes 'on the ground';
- are performed with negligible control over events, notably over who does and does not get guidance;
- may need to be planned into programmes so that appropriate data can be collected as they go along (e.g. from clients before guidance) and, if necessary, before programmes begin - studies are sometimes undermined by the failure to do so;
- are designed from the outset on the assumption that samples are non-random;
- tend to take a 'black box' approach to the actual character of the guidance given, which may be quite heterogeneous;
- often examine public provision from a public-policy perspective.

EXAMPLE: META-ANALYSIS (SPOKANE & OLIVER), 1983; OLIVER & SPOKANE, 1988; WHISTON, SEXTON & LASOFF, 1998)

Meta-analysis reports the combined results of many studies succinctly. Sometimes, the samples of a few similar studies are simply pooled for re-analysis, but in these meta-analyses the characteristics and results of each study form a 'case' in a new analysis.

Each study outcome is converted to an 'effect size' estimate. 'Effect size' is the treated outcome measured in untreated sample standard deviations. This allows one to compare impacts, irrespective of the nature of the outcome variable. But the *value* of an effect depends on what is measured in this way.

The three most important guidance meta-analyses form a series covering 105 studies over a period of approximately fifty years. Spokane & Oliver (1983) was elaborated with some additions as Oliver & Spokane (1988). Whiston *et al.* (1998) considered studies published in the period 1983-1995, taking up where the earlier analyses left off.

Amongst many other things they show that, per session or per hour, one-to-one interventions out-perform group interventions, which out-perform counsellor-free ones. But note that cost-benefit analysis ranks methods differently, and that the criteria of assessment (and value of what is assessed) and types of client are not held constant in this analysis. And, of course, counsellor-free CAGS have advanced in sophistication since most of the studies included were performed.

Sometimes, 'follow-up' studies are conducted in which the counterfactual is merely the starting position of those given guidance[7]. *Ex post facto* studies which make a more serious attempt to estimate the counterfactual are less common. This is because they demand large samples, are difficult to implement and are expensive. Like controlled trials, they make comparisons between people who get guidance and people who do not, or between people who get differing types or amounts of guidance. But as these people are not randomly assigned, *samples are biased with respect to one another* and this bias must be removed. Studies of this kind are, therefore, as good as our understanding of this bias: which is to say, of the factors other than guidance which influence the outcome and which also lead some people to be exposed to guidance whereas others are not.

Armed with this information, we can measure the relevant factors so as to adjust the comparison of treated and untreated samples, providing that there is an adequate overlap - in terms of these factors - between them. This is not the place to enter into the complications and difficulties associated with statistical approaches to this task. Suffice it that, in the past, this was usually attempted by 'regression adjustment' (e.g. Killeen, 1996), but an alternative, called 'propensity score matching', has recently been implemented in the UK (Killeen & White, 2000). This sort of methodology is commonly applied in the evaluation of programmes or funding regimes which have been implemented on a fairly widespread basis. There is generally some attempt to make the treated sample representative of the existing population of people receiving the treatment under consideration. The estimates of the effect of treatment obtained from them are closer to the world of everyday practice than are those which come from small, experimental, demonstration projects.

EXAMPLE: PROPENSITY SCORE MATCHING (KILLEEN & WHITE, 2000)

Note that the method is described here in very simple fashion. Essentially it involves: (a) building a statistical model which discriminates those who participate in guidance; (b) assigning a probability of using guidance from that model to all who did so (not all users are equally probable users); (c) using the same model to assign probabilities of use to those who did not use it; (d) matching non-users to users by their probability of use; and (e) comparing outcomes. This uses the same variables which conventional methods control by regression.

Occasionally in public-policy evaluation research *the counterfactual is defined for a 'super-unit'*. This means thinking of guidance as if it were injected into an organisation or community in the hope that changes in that organisation or community will result. The essential point is that success is not judged against what happens to the individuals who receive guidance, but to what happens to the organisations or communities into which it is introduced. Of course, if done on a randomised basis with adequately sizeable samples of organisations or communities, studies of this sort would belong in the first of our main categories; but, in practice, they are not. The main problem is that evaluators may choose to study 'hard' effects on *rates*, such as a reduction of the course-switching rate in an educational institution or in the local unemployment rate, when it is already known to be difficult to establish corresponding effects at the *individual* level (see later) and when there is no rationale for effects on rates other than through effects on the individuals to whom guidance is given. So, even well-conducted studies taking variables such as local employment rates as outcome measures are unlikely to make convincing demonstrations of effectiveness.

[7] This is called a 'pre-post-test' design, i.e. outcome variables are measured before and after the intervention, and no untreated sample is available for comparison. This is fine in circumstances where the outcome without 'treatment' is certain, but potentially very misleading where it is not.

To summarise, summative guidance evaluation consists mainly of the following:

i. Interpretative evaluation, frequently reliant on small-sample, 'qualitative' participant opinion data. Often concerned also with process issues. But note that intensive studies of this sort can point to the data which should be gathered by the quantitative evaluator and elucidate quantitative findings. This is why it may be included as a stage or element in quantitative evaluation studies.

ii. Quantitative evaluation, which has the central task of *testing against the counterfactual* in order to produce *estimates of the sizes of guidance effects*. There is no necessary connection between detailed methodological approach and detailed purpose, but empirically they are associated, so that there are two main sub-groups:

- Experimental evaluation, consisting of randomised controlled trials and other deliberately contrived experimental studies which, by design or unintentionally, depart from this paradigm. Characteristic of US counselling psychology. Samples usually small. Commonly used to study tightly-defined guidance methods - often innovations.

- *Ex post facto* evaluation of guidance 'as it is' on the ground, whether in the form of a pilot or not. Usually the study of publicly-financed services from a public-policy perspective. Methods by which the counterfactual is specified vary from the over-simplistic, to the use of advanced statistical methods on large samples to compare treated to untreated outcomes. An implicit or explicit aim is to generalise those estimates.

2. Who does it and who funds it?

Most experimental trials of guidance techniques are conducted as a branch of applied psychology and as academic, university-based research. Most of them take place and are published in the USA, by counselling psychologists who specialise in career counselling and by their postgraduate students. The senior figure is sometimes the originator of the technique under evaluation. Such research is conducted on a similar basis, but in much smaller quantity, in Canada, Australia and the UK (e.g. by university teachers of occupational and organisational psychology, or in education departments, and by postgraduate students who may be practitioners).

The disparity in output between the USA and the UK is much larger than would be expected due to their relative size. This is because guidance institutions and their connection to applied psychology have not evolved in the UK in the same way or to the same extent as in the USA. The cause, or consequence, or both is, in summary, that a 'medical model' of innovations in treatments, mostly made in the graduate teaching institutions, tested in trials and disseminated to colleagues and junior practitioners through channels owned and operated by the profession itself, seems more plausible there than here. UK attempts to stimulate evidence-based practice and practitioner research are unlikely to lead to a US-style outcome whilst large institutional differences remain.

Research of this sort is often not specifically funded. In the UK, studies of this sort have only very seldom been conducted by postgraduates supported by the ESRC. The costs of postgraduate study are generally met by the individuals concerned or by their employers. Beyond this, postgraduate costs and academic time are met through general subventions to higher education (HEFCE etc.).

The forms of summative evaluation most widespread in the UK are 'interpretative' studies and quantitative, but *ex post facto*, non-experimental studies. The former are somewhat more associated with 'local' and 'pilot' evaluation, and the latter with 'pilots', 'national pilots' and 'national' evaluation. On occasion, government initiatives have been delivered by local bodies which, contractually, have both undertaken local evaluations and forwarded statistical information to be concerted by a national contractor into a national evaluation, with the intention that the latter should be the basis of impact assessment. Experiences of this model have not been entirely happy.

The contractors which undertake evaluations of each kind include units in universities and colleges (e.g. in the University of London Institute of Education or the Centre for Guidance Studies at the University of Derby) and national research institutes such as the Tavistock Institute, the Policy Studies Institute, the Institute for Employment Studies, the Institute for Employment Research, and the National Institute for Careers Education and Counselling, which may be connected, in some way, to a university, whilst having an independent legal status. These organisations specialise in knowledge of substantive topics and aspects of policy, and in research design and analysis. They commonly conduct qualitative research fieldwork and postal surveys themselves, but do not seek to be providers of mass telephone or personal interviews.

Sub-contracting of fieldwork is therefore common in studies of large samples. This is to market and opinion research organisations such as NFO System Three (formerly PAS), BMRB and MORI, which maintain national 'field forces' of interviewers. Some, like MORI, compete directly for 'social' and hence, upon occasion, guidance survey research contracts[8] of a sort which might otherwise go to organisations like the National Centre for Social Research (formerly SCPR). City consultancy companies (e.g. Coopers and Lybrand) have conducted evaluations on behalf of government departments. Some of these bodies become involved in local evaluation, but the reality is that this is often on too small a scale to be financially viable.

[8] e.g. MORI (1996).

There is a long tail of consultants, small market research organisations, etc., each of which has conducted only one or two guidance evaluation studies, usually from scratch and with little or no awareness of or reference to the accumulated literature. The more 'local' the evaluation is, the more likely it is both that the organisations and consultants performing it are locally-based and that it is not widely disseminated, so that they and their work do not come into general view.

Most national evaluation is funded by government - usually, now, by the DfES, but also the DWP. However, numerous bodies supported by government such as the Learning and Skills Council and the Guidance Council are potential sponsors, acting as a conduit for government funds. At the local level, local authorities and major local bodies such as Local Learning and Skills Councils (or, in the past, the Training and Enterprise Councils) play a similar role.

However, it is very difficult to give a complete account, either nationally or locally, because guidance, conceptualised in a broad fashion - for example, to include careers education - is fairly ubiquitous. Hence, it can appear as *part of,* or at the edges of, numerous other objects of evaluation, such as New Deal (e.g. Winterbotham *et al.*, 2001), Sure Start, the work-related curriculum (Saunders *et al.*, 1997) and so on.

EXAMPLE: EVALUATION OF GUIDANCE AT THE EDGES OF SOMETHING ELSE (DEVINE, REID & THORPE, 1998)

Managed Effective Learning (MELSO) includes some shared elements with careers education and guidance: action planning for learning, use of adults other than teachers, and work experience enhanced to make it 'more meaningful' by emphasising its role in skill development. Devine *et al.* found enhanced work experience to be associated with higher self-reports in skill areas such as working individually, problem solving and taking responsibility. The study showed that certain outcomes are commonly attributed by participants to work experience and are enhanced if work experience is itself enhanced and structured[9].

3. How are research priorities identified?

Priorities are identified for different kinds of evaluation research in very different ways. Innovation in technique is often driven - especially in the USA - by trends in general psychology as applied to the psychology of careers. Other innovations derive from technological change such as the introduction of ICT, or institutional developments in education, and so forth. These provide new objects of evaluation. The usual motive for evaluation is to show they work, or work better than what they replace. However, more complicated questions can also be asked about client-treatment interactions, optimal treatments, cost effectiveness, etc. The main criteria of evaluation employed in such studies represent the concerns of practitioners with a psychological and/or educational approach: the acquisition of knowledge and skills, and attitudinal and other psychological changes.

In the UK, government policy tends to determine both the objects of evaluation (new programmes, or delivery under new funding regimes, in particular) and the criteria of evaluation. The pressure to assess 'hard' outcomes (e.g. effects on educational qualifications employment and wages) is considerable, since guidance institutions are subject to much central government direction and their funding is substantially by the state. The minimal position is that an adequate economic case must be made to the Treasury, since money spent on guidance is regarded as an investment and is money not spent on other priorities, such as health care[10]. Moreover, the case must be made repeatedly, as the socio-economic context changes and political values are contested.

Developments in *how* research is done and, especially, how non-experimental quantitative evaluation is to be undertaken, arise in the policy research, methodological, statistical and econometric literature.

4. How is quality control exercised?

Formalised quality procedures by the organisations conducting research are easier to implement in relation to some aspects of it (e.g. mass fieldwork, coding) than others. The peer-reviewed journals tend to ensure minimum standards in the guidance evaluation research conducted from an academic, applied-psychological perspective. In the UK, policy evaluation research is sometimes subject to less effective peer review, as it is often not published in ways that invite it. But research commissioned by central government often has two 'clients': the operating arm of a government department which is responsible for the programme under evaluation; and an internal, specialist research and evaluation unit. This, coupled with competitive pressures, has the potential to increase adherence to minimum standards.

The greatest difficulties arise when buyers are naïve and researchers are not subject to scrutiny by peers. This most commonly occurs when local evaluation budgets are spent, although national organisations which are 'infrequent buyers' can get into a similar position.

[9] Previous studies and reviews of work experience tending to such conclusions include Conrad & Hedin (1981), Jamieson & Lightfoot, (1982), Sims (1987) and Saunders (1987).

[10] There is virtually no convincing, peer-reviewed evaluation of private-sector guidance, such as outplacement services, against 'hard' outcomes in the UK. Corporate buyers seem either content with the *a priori* case, or able to justify the expenditure in other ways.

5. How is it disseminated?

Much evaluation of spending programmes is not disseminated at all, but used by the funding agency, whether national or local, within a committee or governing body to justify the work done. There is a great waste of potential here, as large studies could be of considerable wider interest and small studies could be aggregated into meta-analyses. However, in the UK some is reported in the form of features in professional magazines such as *Newscheck*, or the magazines of the professional associations such as *Career Guidance Today*. Evaluations of the more structured kind carried out within the academic community are generally published as refereed articles in academic journals.

The main journals in the USA for controlled trial etc. evidence include:

Rehabilitation Counseling Bulletin
Vocational Guidance Quarterly
Journal of Counseling Psychology
Journal of College Student Personnel
Journal of Career Education (NB: 'career education' in the USA includes, but goes well beyond, what we in the UK call 'careers education')
The Counseling Psychologist
Career Development Quarterly
Journal of Employment Counseling
Journal of Counseling and Development
Measurement and Evaluation in Guidance.

In the UK over the last two decades the main sources have been:

British Journal of Guidance and Counselling
DFES Research Reports (formerly *Employment Department*, then *DfEE*, *Research Series*).

And less often, sources such as:

Journal of Occupational and Organizational Psychology (formerly *Journal of Occupational Psychology*),
Journal of Education and Work.

Other non-UK sources include:

Australian Journal of Career Development
International Journal for Educational and Vocational Guidance
International Journal for the Advancement of Counselling.

In some cases, evaluation findings are disseminated through reports targeted at policy-makers and practitioners either in the form of more quickly digestible Briefings (e.g. those produced by the National Institute for Careers Education and Counselling) or as monographs (such as the Occasional Papers produced by the Centre for Guidance Studies at the University of Derby).

6. What sorts of things does it tell us?

UK studies which gather client feedback on guidance tend to paint a picture of satisfaction with services and increased confidence. Many clients feel able to attribute concrete outcomes to their guidance, such as enhanced search behaviour and entry into work, education or training. Of course, younger people, in particular, are more influenced by (Witherspoon, 1995), and tend to *say* (emphasis added) that they are more influenced by, their families, and that guidance institutions are only one of their channels to future opportunities. However, there is an element of oversimplification in such comparisons, as contemporary guidance agencies neither displace the family nor act as the principal gatekeepers of (most) education and work opportunities[11].

Most of the evidence which has been tested against the counterfactual case concerns the 'learning outcomes of guidance'. This is, in part, because learning outcomes occur in the short run, and the potential effect size is big, in comparison to things which are harder to do, such as get employment or enhance wages. For learning outcomes, there is usually no need for lengthy follow-up, samples can be small, and studies are relatively cheap and easy to do.

This is not a drawback: the study of learning outcomes is also consistent with the immediate objective of much guidance practice, which is to help people acquire the knowledge, skills and attitudes which assist *their own* decision making, transition and other career behaviour[12].

For the reasons explained, most of the evidence comes from US controlled-trial studies of specific treatments. Samples tend to be drawn inside the US education system, although important studies of adult samples have occurred (notably, studies of adults on welfare and of other disadvantaged groups). Reviews (e.g. Killeen & Kidd, 1991) show that on a simple 'vote counting' basis, gains are much more frequently reported than null results in each of these categories. Meta-analyses (e.g. Spokane & Oliver, 1983; Oliver & Spokane, 1988; Whiston, Sexton & Lasoff; 1998) tend to confirm this. There is also evidence of effects on associated behaviour, notably on information search.

But from the perspective of policy, learning outcomes are means, not ends. Robust evidence about outcomes corresponding directly to the objectives of policy, especially educational motivation, participation and attainment, employment and wage effects, is scarce. Some of the evidence comes from controlled-trials of specific treatments. But long-term randomised trials of guidance encounter formidable difficulties (e.g. sample attrition; 'contamination' of control samples) and have not been regarded as a priority by relevant

[11] And to the extent that guidance institutions do play a gatekeeper role, the task of estimating their 'effects' on education, training and/or employment becomes more difficult.

[12] There have been many attempts to classify the learning outcomes typically assessed in guidance research, e.g. Killeen & Kidd (1991).

government departments. Thus, evidence often comes from studies of general samples of clients who are compared to eligible but untreated people in a manner statistically adjusted in order to compensate for the absence of random assignment. The difficulty, size, duration and cost of such studies militate against their frequent undertaking.

UK studies have demonstrated positive effects on participation in education and training, both by employed (Killeen & White, 2000) and unemployed adults (Killeen, 1996). If this is the main effect of public provision to adults in the UK, any effects on adult employment and wages are likely to emerge only after quite a long delay - they have yet to be shown. However, studies of intensive guidance interventions with a strong emphasis on supported job search have been conducted in the UK, US and Finland, with mixed results. The best-designed studies have shown the effectiveness of intensive guidance interventions designed to help get people off welfare and into work (e.g. US 'job clubs' - Azrin *et al.*, 1980; 1981; Finnish 'guidance courses' - Vuori & Vesalainen, 1999). There are also hints in the US research that comprehensive guidance programmes in schools may have a small effect on factors such as grades and the perceived value of education as an investment for the future (Lapan, Gysbers & Sun, 1997). UK efforts to use existing general-purpose data sets - notably the Youth Cohort Surveys - to estimate guidance effects have proved disappointing, and those conducting such analyses take the view that they tell us more about the way guidance is distributed than about what it does (e.g. Howieson & Croxford, 1997).

7. What is wrong with it and what should be done?

Perhaps the biggest problem facing guidance evaluation in the UK is that so few people ostensibly responsible for it, either as buyers or sellers of research, are aware of the accumulated body of knowledge. There is quite rapid turnover of the people who commission and undertake guidance evaluation research. This means that it lacks a collective memory. This leads, in turn, to a tendency to assume that effects are easily demonstrated. Thus research designs are too frequently inappropriate to achieve their ambitions.

Three fundamental tasks remain paramount for all forms of evaluation. First, to test the null hypothesis (to establish that there *are* effects); second, to test for significant difference between the effect sizes of alternative treatments and treatments for different kinds of people; and third, to improve specification of the counterfactual.

It might be thought that, from the perspective of public policy, it is only a short step to attach monetary values, both to the effects and to the costs of guidance. But cost-effectiveness and cost-benefit are seldom calculated. Indeed, there is only one widely-available set of calculations of cost-effectiveness (Spokane & Oliver, 1983, in which the cost of guidance, but not the monetary value of effects, was taken into account).

One of the reasons for the absence of cost-benefit calculations is, of course, that those who commission empirical studies do not frame the issue in these terms. It is also likely that, should they do so, it would be difficult to impute monetary values to the so-called 'soft' outcomes commonly measured in guidance evaluation research. Another reason is that, so far as 'hard' outcomes are concerned, such as employment and wage effects, we are really only now beginning to comprehend just how much work must be done to establish reliable estimates of effects upon which cost-benefit calculations should be based.

Too many resources for guidance evaluation are 'wasted', with results generally useable only for the immediate organisation concerned and sometimes not even that. They have nothing to add to the accumulated body of research. We are too dependent for good quality evidence upon the USA, where things are so different that it is, similarly, doubtful how far we should generalise to the UK. In the UK we often say that controlled trials are impractical, when what we really mean is that we do not regard them as important. But given that this is so, one of our aims must be to improve our ability to state the counterfactual and clarify the factors other than guidance which influence the outcome; also those which lead some people to be exposed to guidance whereas others are not. Rather than do poor evaluation, we could prepare to do it well.

References

Austin, M.F. & Grant, T.N. (1981). Interview training for college students disadvantaged in the labor market: comparison of five instructional techniques. *Journal of Counseling Psychology*, 28(1).

Azrin, N.H., Philip, R.A., Thienes-Hontos, P. & Besalel, V.A. (1980). Comparative evaluation of the Job Club program with welfare recipients. *Journal of Vocational Behavior*, 16.

Azrin, N.H., Philip, R.A., Thienes-Hontos, P. & Besalel, V.A. (1981). Follow-up on welfare benefits received by Job Club clients. *Journal of Vocational Behavior*, 18.

Conrad, D. & Hedin, D. (1981). The impact of experiential education on adolescent development. *Child and Youth Services Journal*, 4(3).

Devine, M., Reid, M. & Thorpe, G. (1998). *The Impact of Managed Effect Learning on Key Student Outcomes.* Edinburgh: Scottish Council for Research in Education.

Howieson, C. & Croxford, L. (1997). *Using the YCS to Analyse the Outcomes of Careers Education and Guidance.* Department for Education and Employment Research Series (40). London: HMSO.

References

Imbens, G. & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2).

Jamieson, I. & Lightfoot, M. (1982). *Schools and Industry*. London: Methuen.

Killeen, J. (1996). *Does Guidance Work?: an Evaluation of the Intermediate Outcomes of Gateways to Learning.* Research Studies RS19. London: HMSO.

Killeen, J. & Kidd, J.M. (1991). *Learning Outcomes of Guidance: a Review of Research*. Research Paper No.85. Sheffield: Employment Department.

Killeen, J. & White, M. (2000). *The Impact of Career Guidance on Adult Employed People*. Research Report RR226. Sheffield: Department for Education and Employment. http://www.dfes.gov.uk/research/data/uploadfiles/RB226.doc

Lapan, R., Gysbers, N. & Sun, Y (1997). The impact of more fully implemented guidance programs on the school experiences of high school students: a state-wide evaluation study. *Journal of Counseling and Development*, 75.

Lent, R.W., Larkin, K.C. & Hasegawa, C.S. (1986). Effects of a 'focused interest' career course approach for college students. *Vocational Guidance Quarterly*, 34.

MORI (1996). *Evaluation of Vocational Guidance and Counselling Schemes*. London: MORI.

Oliver, L.W. & Spokane, A.R. (1988). Career intervention outcome: what contributes to client gain? *Journal of Counseling Psychology*, 35(4).

Rolfe, H. (2000). *Improving Responsiveness to the Labour Market among Young People: an Evaluation of Four Pilot Projects*. DfEE Research Report RR 190. Sheffield: DfEE.

Saunders, M. (1987). At work in TVEI: students' perceptions of their work experience. In Gleeson, D. (ed.): *TVEI and Secondary Education: A Critical Appraisal.* Milton Keynes: Open University Press.

Saunders, L., Stoney, S. & Weston P. (1997). The impact of the work-related curriculum on 14- to 16-year-olds. *Journal of Education and Work*, 10(2).

Sims, D. (1987). Work experience in TVEI: student views and reactions - preliminary study. In Hinckley S., Pole, C.J., Sims, D. & Stoney, S.M. (eds.): *The TVEI Experience: Views from Teachers and Students*. Sheffield: Manpower Services Commission.

Spokane, A.R. & Oliver, L.W. (1983). The outcomes of vocational intervention. In Walsh, W.B. & Osipow, S. H.(eds.): *Handbook of Vocational Psychology, Volume 2*. Hillsdale, NJ: Lawrence Erlbaum.

Vuori, J. & Vesalainen, J. (1999). Labour market interventions as predictors of re-employment, job seeking activity and psychological distress among the unemployed. *Journal of Organisational and Occupational Psychology*, 72.

Whiston, S.C., Sexton, T.L. & Lasoff, D.L. (1998). Career-intervention outcome: a replication and extension of Oliver and Spokane. *Journal of Counseling Psychology*, 45(2).

Winterbotham, M. & Adams, L., with Hasluck, C. (2001). *Evaluation of New Deal for Long Term Unemployed People Enhanced National Programme*. ESR82. Sheffield: Employment Service Research and Development.

Witherspoon, S. (1995). *Careers Advice and the Careers Service: the Experiences of Young People*. Employment Department Research Series: Youth Cohort Report 33. Sheffield: Employment Department.